National Research Council Canada
Conseil national de recherches Canada
Institute for Information Technology
Institut de Technologie de l'information
ERB-1058

NRC·CNRC

Comparing Case-based Reasoning Classifiers for Predicting High Risk Software Components

Khaled El Emam, Saida Benlarbi,
and Nishith Goel
September 1999

Canada
NRC 43602

National Research Council Canada
Institute for Information Technology

Conseil national de recherches Canada
Institut de Technologie de l'information

Comparing Case-based Reasoning Classifiers for Predicting High Risk Software Components

Khaled El Emam, Saida Benlarbi,
and Nishith Goel
September 1999

Copyright 1999 by
National Research Council of Canada

Permission is granted to quote short excerpts and to reproduce figures and tables from this report, provided that the source of such material is fully acknowledged.

Comparing Case-Based Reasoning Classifiers for Predicting High Risk Software Components

Khaled El Emam
National Research Council, Canada
Institute for Information Technology
Building M-50, Montreal Road
Ottawa, Ontario
Canada K1A OR6
Khaled.El-Emam@iit.nrc.ca

Saida Benlarbi
Cistel Technology
210 Colonnade Road
Suite 204
Nepean, Ontario
Canada K2E 7L5
Benlarbi@cistel.com

Nishith Goel
Cistel Technology
210 Colonnade Road
Suite 204
Nepean, Ontario
Canada K2E 7L5
ngoel@cistel.com

Abstract

Case-based reasoning (CBR) has been proposed for predicting the risk class of software components. Risky components can be defined as those that are fault-prone, or those that require a high amount of effort to maintain. Thus far evaluative studies of CBR classifiers have been promising, showing that their predictive performance is as good as or better than other types of classifiers. However, a CBR classifier can be instantiated in different ways by varying its parameters, and it is not clear which combination of parameters provides the best performance. In this paper we evaluate the performance of a CBR classifier with different parameters, namely: (a) different distance measures, (b) different standardization techniques, (c) use or non-use of weights, and (d) the number of nearest neighbors to use for the prediction. In total, we compared 30 different CBR classifiers. The study was conducted with a data set from a large real-time system, and the objective was to predict the fault-proneness of its components. Our results indicate that there is no difference in prediction performance when using any combination of parameters. Based on these results, we recommend using a simple CBR classifier with Euclidean distance, z-score standardization, no weighting scheme, and selecting the single nearest neighbor for prediction. The advantage of such a classifier is its intuitive appeal to nonspecialists, and the fact that it performs as well as more complex classifiers.

1 Introduction

Mounting evidence demonstrates that most faults are found in only a few of a system's components [17][42][25][45]. This makes the early identification of faulty components an important research endeavor in that it allows an organization to take mitigating actions, such as focus defect detection activities on components that are likely to be faulty, for example by optimally allocating testing resources [22], or by redesigning components.

Early identification of faulty components is commonly achieved through a binary *quality model* that classifies components into either a *faulty* or *not-faulty* category [7][38]. These binary classification models almost always utilize product metrics as predictor variables.[1] The theoretical basis for developing such binary classifiers has been articulated in [9]. There it is hypothesized that the structural properties of a software component (such as its coupling) have an impact on its cognitive complexity. Cognitive complexity is defined as the mental burden of the individuals who have to deal with the component, for example, the developers, testers, inspectors, and maintainers. High cognitive complexity leads to a component exhibiting undesirable external qualities, such as increased fault-proneness and reduced maintainability.

[1] In some cases aggregates of the product metrics are used, such as domain metrics [30][32] derived from principal components analysis.

A promising modeling technique that has been applied to this quality modeling problem is case-base reasoning (CBR).[2] For example, one empirical evaluation found a CBR classifier to be competitive to a discriminant analysis model [32]. Another study that considered models for predicting the number of faults found a CBR system to have superior predictive performance to an ordinary least squares regression model [18]. Furthermore, it has been argued that CBR is in a class of modeling techniques more suited to the analysis of software engineering data (compared with classical statistical techniques, such as least squares regression) [20].

However, there are many different ways that one can construct a CBR classifier by choosing its parameters. Specifically:

- Which distance measure should be used ?
- Should weights be used or should all independent variables be considered equally important ?
- If the equal importance assumption is invoked, which standardization technique should be used ?
- How many neighbors should be used for making a prediction ?

While one can make some theoretical arguments for selecting the appropriate parameters, it remains of particular importance to evaluate the predictive performance of the different CBR classifiers empirically.[3]

In this paper we evaluate the parameters of a CBR classifier in the context of predicting the risk of software components using *source code* metrics. A high risk component is defined as one that has one or more faults detected during acceptance testing. Therefore, we focus on the prediction of the *fault-proneness* of software components. We use a large data set from an operational real-time system for performing the comparison of CBR classifier parameters with respect to prediction accuracy.

Briefly, our results indicate that there is no difference in predictive performance among all of the CBR classifiers that were considered. These results suggest that the use of a simple CBR classifier that is intuitive to nonspecialists would be preferred since it would achieve similar performance. This preferred classifier would use Euclidean distance, z-score standardization, no weighting scheme, and selecting the single nearest neighbor for prediction.

The paper is organized as follows. Section 2 provides background on the CBR classifiers that we evaluate and the criterion that we use for evaluating predictive performance. In Section 3 our research method is described. This is followed in Section 4 by the results of the study, and a discussion. We conclude the paper in Section 5 with a summary and directions for future research.

2 Background

2.1 Case-Based Reasoning Classifiers

A CBR classifier uses previous "similar" cases as the basis for predicting the risk class of an unseen software component. Previous cases are stored in a *case base*. Similarity (or equivalently, distance) is defined in terms of a set of product metrics. The rationale is that components in the case base that have a similar structure to an unseen case will also exert a similar mental burden on the development staff, and consequently will have a similar fault-proneness.

Khoshgoftaar et al. [32] noted a number of advantages of a CBR classifier. Most notably in our context is that the detailed characterization of the "similar" cases can help one interpret the automatic classification results. In principle, as well, a CBR classifier would provide a straight forward approach for dealing with

---

[2] In non-software domains, at least one previous study that constituted an extensive evaluation of CBR compared with a discriminant analysis model demonstrated the superiority of the CBR classifier (see [2]).

[3] Prediction performance is not the only factor that can be used to evaluate binary classifiers. For example, it has been noted that ease of interpretation of the resultant output and the ability to handle missing data are also important features [6][20]. However, prediction performance is still the feature of a classifier that is of primary importance. For example, if a classifier had an elegant solution to the missing value problem but could not predict high risk components better than a guessing classifier does, then it is doubtful that it will be applied in practice.

missing values. However, in the context of quality prediction using product metrics, there are rarely missing values.[4]

When evaluating the predictive performance of a CBR classifier, one first constructs a *case base* of previous components where the source code metrics and the fault-proneness are known. A different data set is then used as the *test* set. This can be achieved in a number of different ways, such as a holdout sample, cross-validation, bootstrapping, multiple releases, or random subsets. For our study we use random subsets, and this is explained further below.

2.2 Notation

We first present some notation to help explain the different CBR parameters and our evaluation method.

A classification algorithm is used to develop a classifier that assigns a binary value to unseen observations. We restrict ourselves to binary risk classes: High and Low (i.e., the risk of having a fault). Let x_i be the vector of product metrics of the i^{th} component in the test data set, and x_{ij} be the j^{th} product metric measured on the component. Further, let c_k be the vector of product metrics of the k^{th} component in the case base, and let c_{kj} be the j^{th} product metric measured on that component. Furthermore, let the case base be defined by the set $D_X = \{c_1, c_2, \ldots, c_k, \ldots, c_{n-1}, c_n\}$ and the test set of unseen cases be defined as $D_R = \{x_1, x_2, \ldots, x_i, \ldots, x_{p-1}, x_p\}$.

2.3 CBR Parameters

In setting up a CBR classifier, a number of different parameters can be specified that can potentially have a nonnegligible impact on the classifier's prediction accuracy. Specifically:

- Which distance measure should be used ?
- Should weights be used or should all independent variables be considered equally important ?
- If the equal importance assumption is invoked, which standardization technique should be used ?
- How many neighbors should be used for making a prediction ?

Below we describe these parameters and the particular values that we evaluated, including justifications for the selections made.

2.3.1 Distance Function

A multitude of different distance functions can reasonably be used in a CBR classifier. We limit ourselves to the ones that have been actually used in previous software engineering studies, which also happen to be common distance functions in other disciplines. We also limit ourselves to the context where the predictor variables are continuous, since in almost all quality modeling studies this is the case.

Kaufman and Rousseeaw [27] define the Minkowski distance as follows:

$$d_{ik} = \left(\sum_j w_j \left| c_{kj} - x_{ij} \right|^q \right)^{\frac{1}{q}} \qquad \textbf{Eqn. 1}$$

The w_j value is a weight and is discussed below. The most commonly used distance functions for continuous variables are the Euclidean and Manhattan distances. In previous quality modeling studies that used CBR for analysis, both the Euclidean [32] and the Manhattan [18] distance functions were used. We therefore compare these.

---

[4] The exception is when, for example, "include" files are considered.

The Euclidean distance between a component i in the test data set and a component k in the case base is given by[5] (i.e., $q = 2$):

$$Euclidean_{ik} = \sqrt{\sum_j \left(w_j \left(c_{kj} - x_{ij} \right)^2 \right)} \qquad \textbf{Eqn. 2}$$

The Manhattan distance is defined by (i.e., $q = 1$):

$$Manhattan_{ik} = \sum_j \left(w_j \left| c_{kj} - x_{ij} \right| \right) \qquad \textbf{Eqn. 3}$$

A priori, there is no compelling reason to prefer one distance function of another, and therefore, it is prudent to evaluate them empirically.

2.3.2 Weights

Some previous studies with CBR classifiers used weights [32], while others assumed equal weights [18]. The rationale for applying weights is that not all product metrics are equally related to fault-proneness. A weighting method suggested in [36][54], and applied in [32] is to use the estimated parameters from a regression model. Given that our dependent variable is binary, we use the estimated parameters from a logistic regression model (see [24]) as weights.[6,7]

2.3.3 Standardization

Standardization rescales the predictor variables (i.e., the product metrics), and accounts for the fact that not all variables are measured on the same units. We consider four different approaches.

A simulation study in the area of cluster analysis found one type of standardization scheme to be superior in recovering the underlying cluster structure under different conditions including error free data, and data with noise and with outliers [41]. However, the parameters of that simulation are not necessarily reflective of software product metrics data, and therefore this standardization scheme, referred to as Z_5, is evaluated here:[8]

$$Z_5 = \frac{c_{kj} - \min(c_k)}{\max(c_k) - \min(c_k)} \qquad \textbf{Eqn. 4}$$

As will be noted, this has a non-robust denominator in that it will be easily affected by even a single outlier. A more traditional standardization scheme that makes the unit of the variables the sample standard deviation is the z-score:

$$Z - score = \frac{c_{kj} - \frac{1}{n}\sum_k c_{kj}}{\sqrt{\frac{1}{n-1}\sum_k \left(c_{kj} - \frac{1}{n}\sum_k c_{kj} \right)^2}} \qquad \textbf{Eqn. 5}$$

---

[5] Here, the weighted Euclidean distance function is different from [32] whereby the weight is also squared, suggesting that the weight is actually part of the distance rather than weighting the distance. We follow the more common formulation in [27].

[6] The weights are derived from the regression using the case base. Then these are applied to the test set.

[7] It should be noted that there are substantial differences between using the logistic regression parameter estimates to weight the distances, and using the logistic regression parameter estimates directly for prediction. In the former the parameter estimates are augmented with further information from the distance function and therefore they can produce quite different results from the latter. One can consider this as a hybric of a logistic regression and a CBR model.

[8] All standardization equations are defined here in terms of the case base. However, they can also be easily reformulated in terms of the test set by changing the c to an x.

The z-score has been used in previous software engineering studies [32][18].

A more robust approach for standardization is to use the *mean absolute deviation* [55] instead of the standard deviation in the denominator. This is robust because the deviations are not squared, therefore atypical points do not exaggerate it [27]. Robustness is desirable because product metrics data typically have a few extreme values that may exert strong influence on the analysis results (e.g., see [43]).

$$Mean\ Absolute = \frac{c_{kj} - \frac{1}{n}\sum_k c_{kj}}{\frac{1}{n}\sum_k \left| c_{kj} - \frac{1}{n}\sum_k c_{kj} \right|} \qquad \textbf{Eqn. 6}$$

An even more robust measure of dispersion that can be used is the *median absolute deviation* [23], which is:

$$Median\ Absolute = \frac{c_{kj} - med(c_j)}{med\left(\left| c_{kj} - med(c_j) \right|\right)} \qquad \textbf{Eqn. 7}$$

2.3.4 Nearest Neighbor

It can be expected that the number of nearest neighbors that are used as the basis for prediction will have an impact on the prediction performance. Previous software engineering studies with CBR systems [32][18] used the single nearest neighbor as the basis for prediction. We evaluate one, three, and five nearest neighbors.

The CBR classifier that we implemented uses a majority vote to make the prediction in the cases with 3 and 5 nearest neighbors. If there are ties in the k^{th} nearest neighbor, all candidates are included in the vote. This is the approach used in [53].

2.4 Evaluation Criterion

The coefficient that we use for evaluating and comparing the prediction accuracy is Youdon's J coefficient [57], which is used for evaluating diagnostic tests in medical methodology texts [3]. This coefficient addresses problems in current binary classifier evaluation criteria used in the software engineering literature, such as proportion correct accuracy, type I and type II error rates, correctness, completeness, and the Kappa coefficient. The definition of J and its justification are presented in the appendix. Youdon provides estimates of the standard error of J, which can be used to construct confidence intervals. The J coefficient can vary from minus one to plus one, with plus one being perfect accuracy and –1 being the worst accuracy.

3 Research Method

3.1 Data Source

Our data set comes from a large real-time system. It was written by professional programmers in a large commercial organization. This embedded computer application included numerous finite state machines, and interfaces to other kinds of equipment. It was written in a procedural language similar to C. The focus of our study was one subsystem.

The subsystem had a total of 13577 components. Of these, 11012 components were not "include" files, and therefore it was possible to extract the relevant source code product metrics. These latter components amount to more than 10 million lines of code. A total of 8585 components (78%) did not have faults in them, and the remaining 2427 components did. All faults were discovered during acceptance testing.

For the purpose of our study we extracted two random samples to serve as the case base and the test set respectively. Each consisted of 500 observations. The procedure followed was as follows. A simple random sample without replacement of 500 observations was taken from all 2427 observations with faults found in them. Another sample of 500 observations was similarly drawn from the 8585 non-faulty components. For each of the faulty and non-faulty sample, we assigned half the observations to the case base, and the other half to the test data set. Therefore we end up with a case base of 500 observations, 250 of which are faulty and 250 are not faulty. This also applies to the test data set. While the prevalence of faulty components in our case base and test data set do not match the prevalence in the whole system, we did use an evaluative coefficient that is not influenced by prevalence (see the appendix).

3.2 Product Metrics

Below we describe each of the 9 product metrics that were used in our analysis, and provide references to previous studies that have used the same or similar metrics in the past. Since our objective in the current study was to compare CBR classifier parameters, we used metrics that were developed in the past and that have been used in previous studies rather than develop a new or different set of metrics. Furthermore, this set of metrics is collected by the organization with whom the study was conducted, and are currently used for project decision making. In addition, all of these metrics are actively used to evaluate software products during the acquisition of software at Bell Canada [13][39].

The unit about which the metrics were collected is a routine (or procedure). The unit of analysis is a module, which may contain more than one routine.

Metric	Description
Arcs	The total number of arcs found in the control graph. This metric was used in [33].
FanOut	The number of distinct calls to other modules. This is similar to the fan-out metric of Kafura and Henry [26], and has been used in [30][33][29].
Breach	Number of control structure breaches. This represents the number of arc crossings in the control graph that violate the principles of structured programming (e.g., usage of one-entry, one-exit control structures such as loops and conditional statements). This has been used in [29].
Knots	Number of knots. This represents the number of arc crossings in the control graph, and includes crossings that violate the principles of structured programming (breaches of structure).
LOC	Total number of lines of code. This excludes lines in included files, preprocessor directives, and empty lines. Comment lines are included. This value has a Pearson correlation of 0.967 with number of source lines of code. Therefore, it does not matter which one is used. This is a commonly used metric in quality models since size is expected to be an important contributor to fault-proneness.
Entry	The number of entry nodes in the control graph.
Pending	Number of pending nodes. A pending node can be seen as a not attainable path in the module. Even though every possibility of the conditional statements are tested, a pending node cannot be accessed. When programming style conventions require that there be no "dead code", a positive value on this metric indicates that code intended to be functional is not executable [28][29].
Cyclomatic	The cyclomatic complexity number [40].
Global	The number of global variables that are used. This metric was used in [44][33].

Table 1: Product metrics collected.

The above source code metrics were specifically selected since previous work has identified that they are not necessarily redundant [28].[9]

3.3 Dependent Variable

The dependent variable that we used in our study was the detection or non-detection of a fault during acceptance testing. If a fault is detected then the component is considered high risk, and if not then it is considered low risk.

4 Results

4.1 Descriptive Statistics

Table 2 shows the mean and standard deviation of all the product metrics that were considered during our study.

	Case Base		Test Data Set	
Metric	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Arcs	373.8	590.2	355.8	617
FanOut	96.65	148	92.2	144
Breach	15.4	121.5	6.4	37
Knots	152.4	731	100.8	437
LOC	1256	1878.6	1202	1950
Entry	22	31.4	21.5	34
Pending	7.8	21	6.2	14.3
Cyclomatic	148	228	141	236.7
Global	894.5	1442	880	1518

Table 2: Descriptive statistics for the product metrics.

The fact that the standard deviation is much larger than the mean in all cases, and that none of these metrics can go below zero (i.e., they are counts) indicates a heavily skewed distribution. There is a large number of observations with low values on these product metrics, and few with large values. In general, it can also be stated that there is a good concordance between the case base and the test set (some variations would be expected). The biggest discrepancy was in the Breach and Knots variables. Further examination indicated that the case base had a small number of extreme points, hence exaggerating the mean and standard deviations in that set.

4.2 Results from Varying CBR Parameters

The results from varying the CBR parameters are summarized in Table 3, Table 4, and Table 5 for nearest neighbor, nearest three neighbors, and nearest five neighbors respectively. The tables show the J coefficient values.

It would seem from these results that there is a steady progression of improvement as the number of nearest neighbors increases. However, it is more difficult to draw conclusions from these tables on which

[9] Although the use of principal components analysis was a possibility for reducing the number of variables, previous work does indicate that this may not improve the accuracy of classifiers [38].

particular combination of parameters produces the best predictive performance since the differences in the J coefficient are not marked.

	Equal Weight (standardization method)				Weighted
Distance Measure	**Z score**	**Mean Absolute**	**Median Absolute**	**Z_S**	
Euclidean	0.172	0.204	0.208	0.176	0.176
Manhattan	0.208	0.212	0.184	0.182	0.18

Table 3: Evaluation results in terms of the J-coefficient for the different parameters with one nearest neighbor.

	Equal Weight (standardization method)				Weighted
Distance Measure	**Z score**	**Mean Absolute**	**Median Absolute**	**Z_S**	
Euclidean	0.248	0.224	0.216	0.256	0.196
Manhattan	0.224	0.236	0.224	0.284	0.224

Table 4: Evaluation results in terms of the J-coefficient for the different parameters with the three nearest neighbors.

	Equal Weight (standardization method)				Weighted
Distance Measure	**Z score**	**Mean Absolute**	**Median Absolute**	**Z_S**	
Euclidean	0.256	0.28	0.256	0.208	0.288
Manhattan	0.28	0.272	0.292	0.193	0.236

Table 5: Evaluation results in terms of the J-coefficient for the different parameters with the five nearest neighbors.

The charts in Figure 1, Figure 2, and Figure 3 show the 95% confidence intervals for all the classifiers that we considered. The notation that is used is "E" or "M" for either the Euclidean distance or the Manhattan distance, and followed by either the standardization method's name (z-score, mean absolute, median absolute, or Z_S) or that it was weighted.

These charts make clear two interesting conclusions. First, all of the binary classifiers considered here have lower tails for the J values greater than zero, indicating that all perform better than one would expect from a guessing classifier. This is somewhat contrary to the strong conclusions drawn in [38] about the utility of binary classifiers.

The second conclusion that can be drawn from our results, given that all confidence intervals overlap, is that the predictive performance of all the 30 classifiers is essentially the same. There is no compelling reason to prefer a set of parameters in preference to another set of parameters when implementing a CBR classifier.

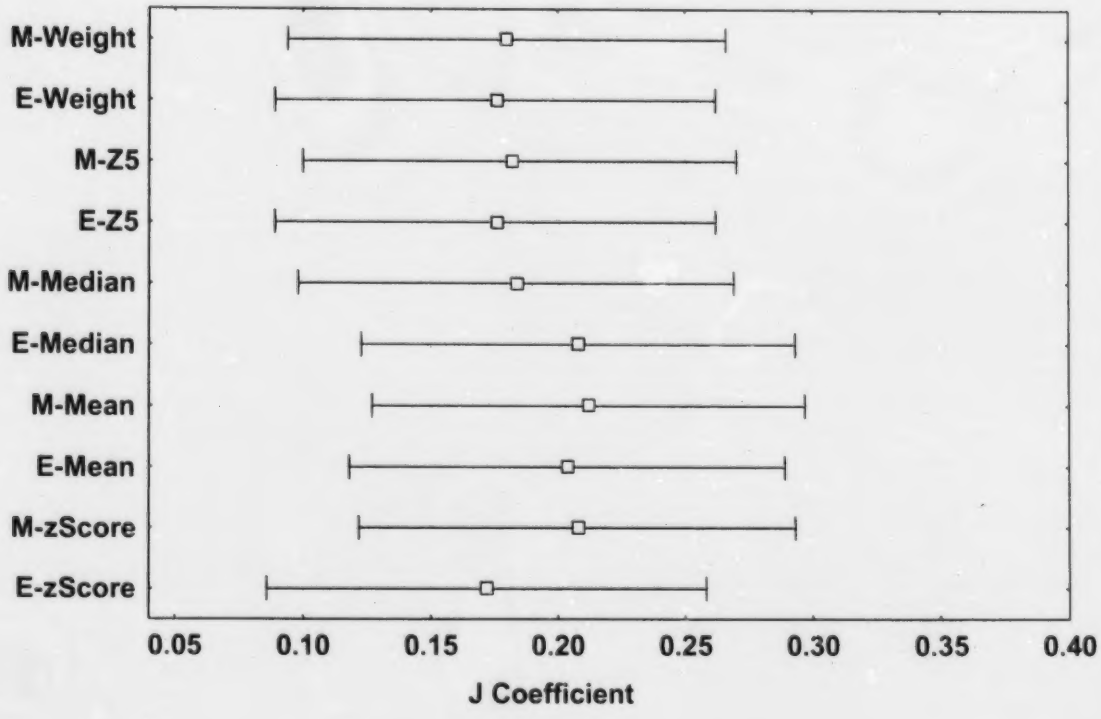

Figure 1: 95% confidence intervals for the J coefficient with the nearest neighbor CBR classifiers.

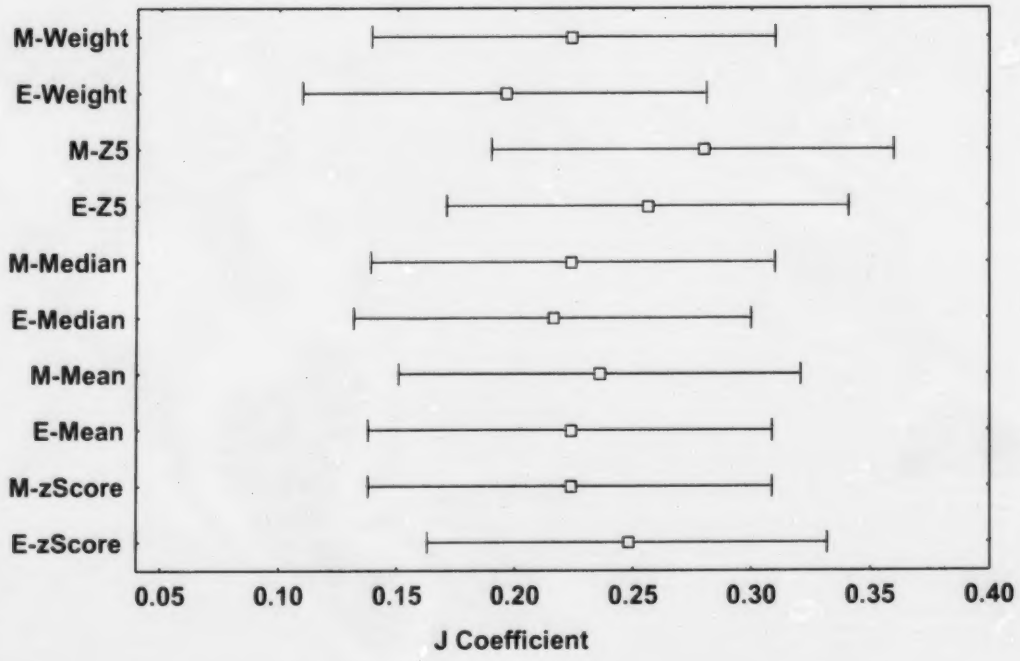

Figure 2: 95% confidence intervals for the J coefficient with the three nearest neighbor CBR classifiers.

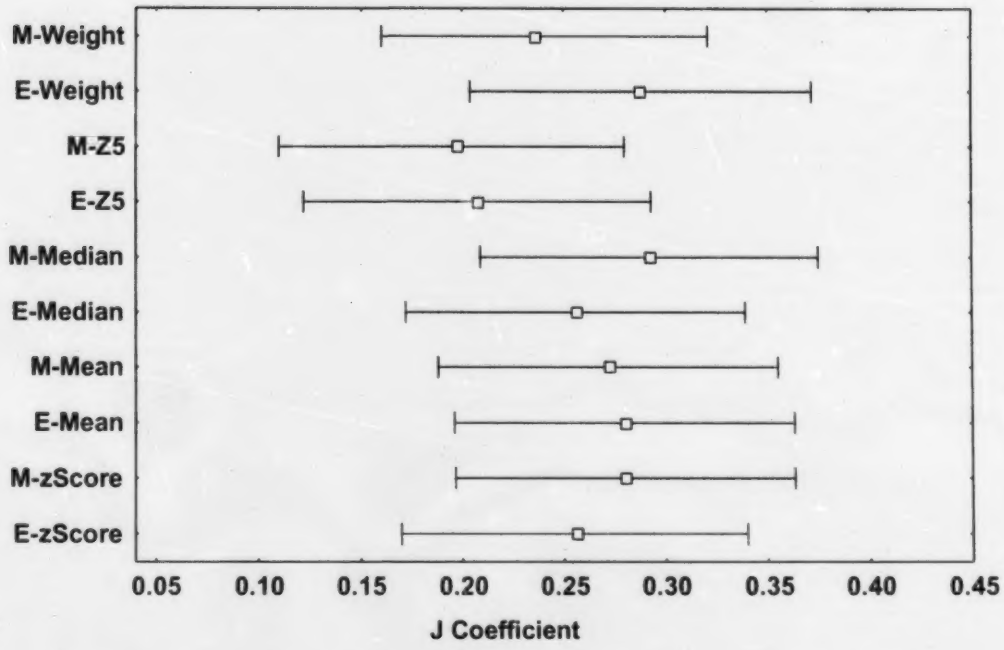

Figure 3: 95% confidence intervals for the J coefficient with the five nearest neighbor CBR classifiers.

4.3 Discussion

The results that we obtained would support the recommendation that, when one is constructing a CBR classifier for quality modeling, it would be preferred to use the simplest CBR classifier. The reason for this being that their predictive performance is the same. The simplest CBR classifier is one that uses a Euclidean distance and the z-score for standardization, and that selects the nearest single neighbor for making the prediction. This is exactly the classifier that was used in an earlier study by Khoshgoftaar et al. [32]. The reason for using the simplest one is that it is quite easy to implement and is intuitive in its mechanisms, requiring no special knowledge to understand.

Two further methodological notes that ought to be emphasized are: (a) proper accuracy coefficients should be used that are not affected by prevalence, such as the J coefficient, and (b) it would be of value for future research to consider the confidence intervals when comparing binary classifiers. By just considering the obtained J values, one would have been tempted to conclude that a more complex classifier (to build and to understand) has the best predictive performance. Where, in fact, the evidence when one considers sampling variability does not support that conclusion.

5 Conclusions

The objective of this paper was to evaluate the predictive performance of case-based reasoning classifiers with different parameters. An empirical evaluation was conducted using a large data set from a large real-time system. Our results indicate that there is no difference among all of the different CBR classifiers. We therefore recommended the use of a simple CBR classifier due to its intuitive appeal to nonspecialists and the fact that our evidence shows it performs as well as other classifiers.

Furthermore, our methodology re-enforces the need to consider prevalence independent measures of prediction accuracy and the use of confidence intervals in the comparison of binary classifiers.

It would be informative for future research to first focus on confirmatory studies of our results, and secondly to compare the predictive performance of other classification techniques with this simple CBR classifier to determine if any of them really are superior. While such comparisons may not unearth a truly "best" classifier across all data sets, it may help discount classifiers that do not seem to add much prediction performance value. Furthermore, additional research on the interpretation of the J coefficient would be beneficial for future studies that are not comparative in nature.

6 Appendix: A Coefficient for Evaluating Binary Classifiers

The purpose of this appendix is to show that current criteria that are used in software engineering to evaluate the accuracy of binary classifiers have some important deficiencies. The deficiencies surface when one is either evaluating a binary classifier on a single data set with the hope that the accuracy results will generalize to real projects, and when one is comparing binary classifiers on the same data set. We then propose a new coefficient that addresses these deficiencies. This new coefficient was originally proposed in medicine to evaluate the accuracy of diagnostic tests, and is used in our study.

6.1 Definitions and Notation

A classification algorithm is used to develop a classifier that assigns a risk class to unseen observations. We restrict ourselves to binary risk classes: High and Low. If we let our case base be $D_X = \{c_1, c_2, \ldots, c_k, \ldots, c_{n-1}, c_n\}$, and let R be the space of unseen observations, then a classification algorithm maps a given data set D into a classifier C, and a classifier C maps an unseen instance $r \in R$ to a binary value $y \in$ {High, Low}. We use the notation Z(D,r) to denote the binary value assigned to an unseen case r by the classifier built by the algorithm Z using data set D.

D_R is a new data set that is obtained from a real project that wishes to use the classifier after it has been evaluated. The analyst does not have access to D_R during the evaluation[10], but hopes that the results of the evaluation using D_X are indicative of the performance of the classifier when it is used in the future on D_R. Therefore, we make a distinction between the *evaluation* phase of using a classifier, and the *application* phase.

An intuitively obvious way to evaluate the overall "goodness" of a classifier is to calculate the proportion of correct classifications. The proportion correct has been used in a number of previous studies to evaluate classifiers [1][38][49].

One commonly used sampling approach for estimating the accuracy of a classifier on unseen cases is to use a holdout sample. With the holdout sample method, accuracy is calculated as:

$$Accuracy = A = \frac{1}{N} \times \sum_{(r_i, y_i) \in D_h} \lambda(Z(D_X, r_i), y_i) \qquad \textbf{Eqn. 1}$$

where D_h is the hold-out sample (the test set in our study), $D_h \cap D_X \cap D_R = \Theta$, $|D_h| = N$, and $\lambda(i, j) = 1$ if $i = j$ and zero otherwise.

Table 6 shows the notation in obtained frequencies when a classifier is used to predict the class of unseen observations in a confusion matrix.

---

[10] Or if s/he does, the values of y in D_R are unknown.

		Predicted Risk Class for D_h		
		Low	High	
Real Risk Class in D_h	Low	n_{11}	n_{12}	$N_{1\cdot}$
	High	n_{21}	n_{22}	$N_{2\cdot}$
		$N_{\cdot 1}$	$N_{\cdot 2}$	N

Table 6: Notation for a confusion matrix.

If the obtained accuracy on D_h is high enough, the analyst would then apply the classifier on real projects (i.e., on data set D_R).

Such a confusion matrix also appears frequently in the medical sciences in the context of evaluating diagnostic tests (e.g., see [19]). Two important parameters have been defined on such a matrix that will be used for our exposition, namely sensitivity and specificity.[11]

The *sensitivity* of a classifier is defined as:

$$s = \frac{n_{22}}{n_{21} + n_{22}} \qquad \textbf{Eqn. 2}$$

This is the proportion of high risk components that have been correctly classified as high risk components.

The *specificity* of a classifier is defined as:

$$f = \frac{n_{11}}{n_{11} + n_{12}} \qquad \textbf{Eqn. 3}$$

This is the proportion of low risk components that have been correctly classified as low risk components.

Ideally, both the sensitivity and specificity should be high. A low sensitivity means that there are many low risk components that are classified as high risk. Therefore, the organization would be wasting resources reinspecting or focusing additional testing effort on these components. A low specificity means that there are many high risk components that are classified as low risk. Therefore, the organization would be passing high risk components to subsequent phases or delivering them to the customer. In both cases the consequences may be expensive field failures or costly defect correction later in the life cycle.

Finally, the hold-out prevalence[12] of high risk components is defined as:

$$p_h = \frac{n_{21} + n_{22}}{N} \qquad \textbf{Eqn. 4}$$

This is the proportion of high risk components in the hold-out sample (i.e., the prevalence of "High" risk components in the test sample D_h that was used as the basis for evaluating the classifier).

---

[11] The terms sensitivity and specificity were originally used by Yerushalmy [56] in the context of comparing the reading of X-rays, and have been used since.

[12] This is also sometimes referred to as the "base rate".

Furthermore, let π be the (unknown) proportion of observations in D_R that have a "High" risk class, and consequently $1-\pi$ is the proportion of observations with a "Low" risk class. We will refer to π as *actual prevalence* (i.e., the prevalence of high risk classes in contexts where the classifier is actually going to be used for making decisions).

6.2 Generalization and Comparison

Software engineering studies that evaluate binary classifiers tend to be of two kinds:

- Evaluating the accuracy of a classifier on some data set where the actual risk classes of the components are known (i.e., build the classifier on D_l and evaluate it on D_h). The results of this evaluation are assumed to be a reflection of how well the classifier will perform when used on D_R (i.e., the evaluation results will generalize).
- The accuracy of two or more classifiers are compared on the same data set. The classifier that has the largest accuracy is assumed to be better. It is further assumed that the same high accuracy of the best classifier will be obtained when it is used on D_R.

It is therefore necessary that studies of classifiers use accuracy measures that will not be conducive to misleading conclusions under both of the above contexts.

6.3 Defining a "Good" Classifier

Intuitively appealing and necessary conditions for a binary classifier to be considered "good" during evaluation are twofold:

- It must perform better than would be expected by simple guessing (i.e., assigning unseen observations to classes at random with prior probabilities of 0.5)
- It must do better than always assigning observations to the most frequent class (i.e., if $p_h > 0.5$ then always classify as "High" risk, and if $p_h < 0.5$ then always classify as "Low" risk).

With a binary classification, one would expect to obtain 50% proportion correct accuracy by guessing with no prior information[13]. As the accuracy increases above 50%, the more confidence one gains about the usefulness of the classifier for prediction[14]. In many cases, software engineering researchers take 50% as an implicit minimal accuracy for a binary classifier[15]. This threshold has also been made explicit in the past [38]. Therefore, classifiers that have accuracies above 50% are considered to be "good".

In many cases it is possible to obtain rather good proportion correct accuracies by simply classifying observations in the most frequent category. This is particularly true when there are few "High" risk components and therefore classfiying everything as "Low" risk would still give a good proportion correct accuracy result.

---

[13] With no prior information means assigning observations to either of the Low or High risk classes with a probability of 0.5.

[14] Most research in quality modeling does not take misclassification costs into account (except for example [34]). This is partly due to the difficulty in obtaining data on the cost of misclassification. Improved estimates about the usefulness of classifiers could be made if such costs were available.

[15] No evaluative judgement is made here about whether this is the best accuracy value to use. Only that it is commonly used, indicating acceptance amongst researchers as a reasonable threshold for binary classifiers.

		Predicted Risk Class		
		Low	High	
Real Risk Class	Low	80	0	80
	High	5	0	5
		85	0	85

Table 7: An example of a classifier that always predicts all risk classes as "Low".

Table 7 gives an example of such a classifier. Here the classifier always predicts "Low" risk, however its proportion correct accuracy is a respectable 94%. Few would argue that such a classifier is very useful, despite its high proportion correct accuracy.

While this specific problem has not been articulated explicitly in the software engineering literature, in other areas where validation coefficients for nominal scales were considered, corrections for this kind of "accidental" accuracy have been proposed [5].

We therefore expect that any criterion for evaluating and comparing binary classifiers to account for both of the above, otherwise misleading conclusions may ensue.

6.4 Software Engineering Contexts for Evaluating Classifiers

In this subsection we show that in software engineering studies, it is common that $p_h \neq \pi$. This has important implications on the types of coefficients that are used to evaluate binary classifiers.

It is frequent in software engineering studies that the dependent variable is not binary in nature, but is dichotomized by the researchers. For example, the dependent variable may be the number of field defects found. This would then be dichotomized into "High" and "Low" categories.[16]

The dichotomization may be based on expert opinion (for example, as in [31]), or may be based on the data itself. When dichotomization is based on the data itself, one of the following options is frequently followed:

- The authors perform a median split of the dependent variable, and then use a leave-one-out cross-validation approach to estimate the accuracy values. For example, see [1][4]. The consequence of this is that $p_h \approx 0.5$. It is unrealistic to assume that in D_R exactly half of the components will be high risk, and therefore $p_h \neq \pi$. In other studies a sample is constructed to ensure equal numbers of high and low risk components [8].
- In other studies the median split is performed on the training data set, and the *same* cutoff value is used on the hold-out sample which can be a different system within the same organization. As is clear in [16], median splits on one system do not result in the same prevalence for another system even within the same organization, i.e., this does not necessarily result in a $\pi = 0.5$.
- Some authors perform splits on quartiles, such as [32] where splits were performed on the third quartile. In addition to the fact that quartile splits do not guarantee any more than do median splits that $p_h = \pi$, in this particular study the quartile split was performed on the hold-out sample as well (i.e., the cutoff point was different for the training and the hold-out samples rather than being the same). This will ensure that $\pi = 0.25$ only if the dependent variable in D_R is already known, which is not the case.[17]

---

[16] Where dichotomization is not performed and the prediction is made of the number of defects found directly, then this is outside the scope of our paper since we are only interested in binary classifiers.

[17] If the dependent variable values are known in D_R then there is no point in building a classifier.

- Student programs are used and the holdout sample is selected randomly. However, notwithstanding any other deficiencies in using student programs, it is not clear to what extent p_h is representative of any realistic π in an application context, irrespective of which of the above techniques is used for splitting.

While the above scenarios do not characterize all studies that evaluate binary classifiers, they still do represent a significant proportion of the studies that are conducted that any evaluation criterion must be able to accommodate them.

6.5 Evaluative Measures in Software Engineering

A number of different measures are used to evaluate binary classifiers in software engineering research. Some of these are descriptive and some are inferential. We review these briefly below.

6.5.1 Chi-Square Test

Some software engineering authors have used a chi-square inferential test for the 2x2 table as the basis for determining predictive validity [4][1][38][50]. It has also been suggested that classifiers that do not pass the chi-square test of significance should be discarded [38].

The current criterion is to first calculate a χ^2 value. This value compares the observed confusion matrix with the expected confusion matrix under the null hypothesis of independence between the column and row variables. If the value of an obtained χ^2 is larger than a critical value then it is claimed that the classifier has predictive validity. The critical value can be obtained from tables of the chi-square distribution with one degree of freedom for a given alpha level.

It has been noted that this test evaluates the capability of a classifier to predict future behavior from past behavior, or the ability of a classifier to discriminate between high risk and low risk software components [38]. Furthermore, it has been such that a classifier that does not meet the criterion of predictive validity using this test is rejected from further consideration [38].

6.5.2 Sensitivity and Specificity

Some authors, such as [1], report the sensitivity and specificity values directly.

6.5.3 Proportion Correct

Most authors will report the proportion correct value, for example [1][38][49] (also called correctness in [46][47]). This is an intuitively appealing measure of prediction performance since it is easy to interpret.

6.5.4 Type I and Type II Misclassifications

These two evaluative measures are used in a number of different studies, such as [29][31][35][32]. The Type I misclassification rate is $1-f$, and the Type II misclassification rate is $1-s$.

6.5.5 Completeness and Correctness

Correctness and completeness are also frequently used measures for evaluating binary classifiers. Completeness, as defined in [7][46], is equal to the senstivity. Correctness is the true positive rate [7][9][10][11] (also called consistency in [46] [47]), and is defined as:

$$Corr = \frac{n_{22}}{N_{+2}} \qquad \text{Eqn. 5}$$

The logic of considering correctness is that it gives an indication of how many of the "High" risk predictions made by the classifier are actually correct. If correctness is low then there will be extensive wasted effort since many of the components identified by the classifier as high risk will in actuality be low risk.

6.5.6 Kappa

In some recent studies the Kappa coefficient has been used as a measure of prediction performance of binary classifiers [9][11]. The Kappa coefficient was originally devised by Cohen to evaluate inter-observer agreement in their classifications on a nominal scale [14][18]. It is defined as follows:

$$\kappa = \frac{\frac{n_{11}+n_{22}}{N} - \sum_{i=1}^{2}\frac{N_{+i}N_{i+}}{N^2}}{1-\sum_{i=1}^{2}\frac{N_{+i}N_{i+}}{N^2}} \qquad \text{Eqn. 6}$$

In the above formulation, $\frac{n_{11}+n_{22}}{N}$ is the obtained "agreement" between the actual risk class and the predicted risk class, and is equal to the proportion correct coefficient. Also, $\sum_{i=1}^{2}\frac{N_{+i}N_{i+}}{N^2}$ is the "agreement" that would be expected by chance if the classifier was assigning risk classes at random with probabilities equal to the column proportions. Chance "agreement" assumes that the rows and columns are independent.

The definition of Kappa is then the "agreement" obtained beyond that which would be expected by chance compared to the maximum possible "agreement" that could be obtained.

6.5.7 Summary

In addition to the fact that different authors use different measures, they also frequently report *multiple* measures in a single study. While multiple measures for a single study may be useful for getting an overall intuitive picture of prediction performance, they also present a difficulty in drawing conclusions. For example, if in a comparison study the classifier U was found to better than classifier W on one measure, but the opposite was true on another measure, drawing objective conclusions becomes more difficult. To illustrate this point we refer, for example, to the results in [32]. Here the authors were comparing nonparametric discriminant analysis with case-based reasoning classification. The Type I misclassification rate was lower for discriminant analysis, but the Type II misclassification rate and the overall proportion correct accuracy were lower for case-based reasoning. From these results it is not clear which modeling technique is superior.

Therefore, ideally, there ought to be a *single* evaluative measure that can be used for evaluation and comparison, which in turn would lead to consistently objective conclusions. Below we highlight deficiencies in most of the above measures individually. We divide this into two parts, the inferential measure, namely the chi-square test, and the descriptive measures. This is then followed by a proposed new single measure that addresses these deficiencies.

6.6 An Appraisal of the Chi-Square Test

The chi-square testing approach has two disadvantages:

- This test, when used in the context of evaluating classifiers, does not provide useful information about the value of the classifier. First, when the null hypothesis is rejected, this does not signify that a classifier is necessarily "good". Second, when the null hypothesis is not rejected, this does not necessarily signify that the classifier is "bad".
- In practice, the test is not used alone but in conjunction with another criterion: classification accuracy. Therefore, the real decision process consists of two criteria. The Type I error rate (in a

[18] Cohen also defined a weighted version of Kappa [15], but the characteristics that are of interest to us are the same for Kappa and its weighted version.

statistical sense) of the whole decision process is conservative and results in the analyst actually operating at a lower alpha level than s/he stipulates. This consequently results in loss of power (i.e., inability to identify "good" predictors).

Here we show that the chi-square approach does not actually provide useful information *the way it is applied in practice*. It is shown below that a classifier that *does* pass the chi-square criterion should not necessarily be accepted as "good". Conversely, if the chi-square criterion can identify classifiers that should be rejected (i.e., they do not pass the chi-square criterion), then perhaps it still has some value. However, it also shown that this criterion does reject classifiers that are "good".

The chi-square test is non-directional. This means that association on the left diagonal is as good as association on the right diagonal of the confusion matrix[19]. For example, the confusion matrix in Figure 4 shows a classifier that has a high chi-square value (approximately 15 after a correction for continuity) that is statistically significant at an alpha level of 0.05. The confusion matrix in Figure 5 gives exactly the same result in terms of chi-square. It is clear that the classifier of Figure 4 has greater accuracy (approximately 74%) than that the one in Figure 5 (approximately 26%). Therefore, the chi-square criterion by itself does not tell us whether the classifier is "good" or not.

		Predicted Class: LOW	Predicted Class: HIGH	
Real Class	LOW	25	10	35
	HIGH	10	31	41
		35	41	76

Figure 4: Confusion matrix with significant chi-square and high concordance.

		Predicted Class: LOW	Predicted Class: HIGH	
Real Class	LOW	10	25	35
	HIGH	31	10	41
		41	35	76

Figure 5: Confusion matrix with significant chi-square and low concordance.

Proponents of the chi-square testing approach may argue that the test is not used in a blind fashion. If patterns such as those shown above are obtained, and the test is statistically significant, then it is obvious whether this is "good" or not because of high or low concordance between the actual and predicted values.

However, it is plausible that even though the chi-square test is significant, it is not possible to tell whether this is because of high or low concordance between the actual and predicted values. Consider the confusion matrix in Figure 6. The chi-square test here is statistically significant at an alpha level of 0.05 (with a χ^2 value of 4.96). Exactly half of the classifications are on the right diagonal, and half on the left diagonal, giving an accuracy of 50%.

		Predicted Class: LOW	Predicted Class: HIGH	
Real Class	LOW	19	37	56
	HIGH	1	19	20
		20	56	76

Figure 6: Confusion matrix with unclear concordance.

[19] In all examples we use Pearson's chi-square statistic with Yates' continuity correction. Furthermore, all examples use N's that are common in the software engineering literature (for example, see [1][4].

Let's say that the criterion is changed so that the accuracy has to be greater than 50% for a classifier to be claimed to have predictive validity. Therefore, the whole decision process is actually a two stage one:

- determine whether the chi-square value is statistically significant, then
- if it is, determine whether the accuracy is greater than 50%

If both of the above criteria are satisfied, then it can be claimed that the classifier has predictive validity. In such a case, the null hypothesis that is being tested is not that of independence, but of *independence and accuracy less than or equal to 50%.*

However, this approach would lead to a test that is conservative. This means that the Type I error rate (the probability of rejecting the null hypothesis when it is true) would be lower than the nominal alpha rate. First we illustrate this. We conducted a Monte Carlo simulation to determine the actual Type I error rate if the above decision process was implemented. We chose observation sizes (NOBS) ranging from 10 to 100 in increments of 10. For each number of observations we chose the number of actual Low risk components to be $x \times \frac{NOBS}{10}$ where x was varied from 1 to 5. The classifier we implemented was a guessing classifier that would assign observations to each of the two risk classes with a probability of 0.5. The above scheme for evaluating predictive validity was implemented for an alpha level of 0.1 and 0.05. For each combination of NOBS, x, and alpha value we performed 5000 iterations. Under this simulation we would hope that the proportion where the null hypothesis of *independence and accuracy less than or equal to 50%* to be rejected at the same rate as the alpha level.

The results are shown in Table 8 for an alpha level of 0.1, and Table 9 for an alpha level of 0.05. It is clear from these tables that the null hypothesis is rejected at much smaller proportions than the nominal alpha level. In general, it is rejected half the number of times that you would expect. This means that this approach is conservative, and assigns a lower risk to the test than one was willing to accept. In some cases, it is rejected ten times less. By implication, this also means that the power of this test is reduced.

NOBS/x	1	2	3	4	5
10	0.010	0.039	0.088	0.070	0.059
20	0.056	0.051	0.059	0.066	0.057
30	0.064	0.061	0.059	0.060	0.052
40	0.049	0.046	0.052	0.055	0.044
50	0.045	0.057	0.050	0.049	0.063
60	0.047	0.058	0.051	0.053	0.044
70	0.054	0.051	0.052	0.053	0.060
80	0.044	0.051	0.051	0.055	0.046
90	0.051	0.056	0.049	0.047	0.052
100	0.047	0.050	0.054	0.050	0.045

Table 8: Proportion of times the null hypothesis is rejected for alpha equal to 0.1.

NOBS/x	1	2	3	4	5
10	0.010	0.012	0.032	0.021	0.034
20	0.010	0.029	0.020	0.030	0.017
30	0.013	0.026	0.028	0.026	0.022
40	0.029	0.021	0.028	0.024	0.022
50	0.025	0.030	0.023	0.026	0.038
60	0.024	0.030	0.028	0.031	0.026
70	0.028	0.024	0.025	0.025	0.024
80	0.022	0.025	0.022	0.025	0.028
90	0.025	0.027	0.024	0.025	0.020
100	0.021	0.028	0.027	0.028	0.027

Table 9: Proportion of times the null hypothesis is rejected for alpha equal to 0.05.

The reason is that the chi-square test would normally reject the null hypothesis of no association a proportion of times that is almost equal to the nominal alpha rate. However, because of the decision to only reject the overall null hypothesis when accuracy is greater than 50%, this means that the null hypothesis would not be rejected in some of the cases where the chi-square test rejects the independence (sub-)hypothesis. The simulation shows the extent to which this would happen.

The above discussion illustrates that when the chi-square test gives a statistically significant result, this by itself does not provide information about whether the classifier is "good" or not. When combined with another criterion, the whole decision process has a rather conservative Type I error rate.

However, when the chi-square test is *not* significant, does it provide useful information ? Can it be used to eliminate classifiers that have no predictive power ? If this is the case, then the test still has merit. Unfortunately, it can be shown that the chi-square criterion by itself would eliminate classifiers that are "good" by the accepted criterion of accuracy greater than 50%.

Consider the confusion matrix in Figure 7. This has a chi-square value that is *not* statistically significant, even though this classifier has an accuracy of 75%, which is quite large. The probability of getting 75% correct classifications for 100 unseen cases is less than 0.0001 for a classifier that guesses at 0.5 probabilities (this can be calculated using a Monte Carlo simulation), which *is* statistically significant[20].

		Predicted Class		
		LOW	HIGH	
Real Class	LOW	70	20	90
	HIGH	5	5	10
		75	25	100

Figure 7: Confusion matrix for a "good" classifier that provides a chi-square test that is not statistically significant.

6.7 An Appraisal of the Descriptive Measures

In this subsection we provide a critical appraisal of the descriptive measures of accuracy. We first present a series of examples that we will use throughout to illustrate our points.

---

[20] The reason this happens is because the chi-square test does not assume under the null hypothesis of independence that a classifier classifies randomly at probabilities of 0.5. Expected frequencies are calculated from the observed marginal totals. In the example, if a random classifier assigns cases to "Low" with 0.75 probability and "High" with 0.25 probability, it would be expect to have an accuracy of 70% in the long run. Therefore, a different threshold is effectively used compared to the one that is assumed for evaluating how "good" a classifier is. Actually, the probability of getting 75 correct classifications or more for the "Real Class" marginal totals in Figure 7 using a guessing classifier that classifies to the Low risk category at a 0.75 probability and the High risk category at a 0.25 probability is approximately 0.137 (this value can calculated using a Monte Carlo run). This is not statistically significant.

6.7.1 Definition of Examples

We define two sets of examples, the first are used to illustrate the generalizability problems, and the second to demonstrate the comparison problems.

Example No.	f	s	p_h	π
1	0.9	0.7	0.9	0.2
2	0.9	0.7	0.2	0.9
3	0.95	0.1	0.5	0.05

Table 10: Examples used to illustrate generalizability problems.

Table 10 shows the examples to illustrate generalizability problems. Example 1 is a classifier with a specificity of 0.9 and a sensitivity of 0.7. The hold-out sample has a prevalence of 0.9 and the actual data set has a prevalence of 0.2 (i.e., the actual data set's prevalence is much smaller than the hold-out sample). The second example is for the same classifier except that the prevalances have been switched. The third example is a classifier that can classify "Low" risk components very well, with a specificity of 0.95, but that is rather bad at classifying "High" risk components. The hold-out sample has 50% of its components high risk, but the actual sample has very few "High" risk components (only 5%).

Example No.	f	s	p_h
4	0.8	0	0.2
5	0.6	0.8	0.2

Table 11: Examples used to illustrate comparison problems.

Table 11 shows examples of two classifiers to be compared. In both cases the hold-out prevalence is the same since we assume that the classifiers are compared on the same data set. In example 4, the classifier classifies all components in the "Low" risk category, hence its senstivity is always zero. The fifth example shows a more balanced classifier.

In all our examples we take it for granted that sensitivity and specificity estimated on the hold-out sample will be stable on a real project data set. If this assumption is not tenable than there is no value to any evaluative study of quality models since then one can never generalize from evaluative studies to actual practice.

6.7.2 Proportion Correct Accuracy

Below we show that the use of proportion correct classifications as a measure of accuracy, and consequently any inferential test based on it (such as that proposed in [16]), can produce quite erroneous results in the context of identifying risky software components. The reason is that A has a strong dependence on the prevalence p_h. We will first show this dependency and then illustrate its detrimental effects with reference to the examples.

We can make the following definitions:

$$n_{22} = s \times p_h \times N \qquad \textbf{Eqn. 7}$$

$$n_{11} = f \times (1 - p_h) \times N \qquad \textbf{Eqn. 8}$$

And therefore

$$A = \frac{(s \times p_h \times N) + (f \times (1 - p_h) \times N)}{N} = (s \times p_h) + (f \times (1 - p_h)) \qquad \textbf{Eqn. 9}$$

Now, consider that a classifier has been developed and it has been evaluated on the holdout sample D_h. From this evaluation, the proportion correct accuracy using D_h, say A_h, has been calculated. Let's say that the value A_h was sufficiently high, and therefore now we want to use it on a real project to make predictions. We expect that the accuracy A_h obtained during the evaluation will be reflective of the accuracy of the predictions that are made on the data set D_R, such that if we were able to calculate A_R, the accuracy on the real project, we would have $A_R \approx A_h$.

In fact, this expectation is quite misguided as, by definition, it would only occur under some rather restrictive conditions.

Let's consider our first example in Table 10. This was for a classifier with $f = 0.9$ and $s = 0.7$ respectively. For this example, we can calculate $A_h = 0.72$ and $A_R = 0.86$. Thus, during the evaluation on data set D_h we have substantially underestimated the proportion correct accuracy of our classifier. A different situation would occur if we reverse the prevalences as in example 2 in Table 10. Then we have $A_h = 0.86$ and $A_R = 0.72$. In such a case, the evaluation on data set D_h overestimates the accuracy of the classifier when it is used in a real project. The only condition when the two accuracies are the same is when $p_h = \pi$ (i.e., the evaluation data set has the same prevalence of high risk components as actual projects).

The reason is that this classifier does a good job of classifying low risk components, and less proficient job at classifying high risk components. Therefore, if the sample has more low risk components, the proportion correct accuracy will be better. But, if the sample has more high risk components, the proportion correct accuracy will deteriorate.

It is quite easy to conclude from this that the proportion correct accuracy is not a reflection of the performance of a classifier on a project with different prevalences from the evaluation data set.[21] This problem is further exacberated because some researchers artificially construct their evaluation data sets so that 50% of the observations are in each class. Example 3 in Table 10 is an extreme case where the classifier does not do a good job in identifying "High" risk components. In such a case, $A_h = 0.525$ and $A_R = 0.91$. This indicates that on the hold-out sample with 50% "High" risk components the classifier performs barely better than our 50% acceptable accuracy, but on the real project data set it has a remarkably high accuracy. Therefore, making the reasonable assumption that the prevalence of high risk components on real systems is considerably less than 50%, it is likely that many past evaluation studies underestimated the proportion correct accuracy of their classifiers through the practice of creating artificial samples, and that many useful classifiers were discarded.

It is also easy to conclude that efforts to construct null hypothesis tests for the proportion correct accuracy by comparing obtained values with the "guessing" null distribution (such as [16]) do not solve the problem since they still rely on the proportion correct accuracy as the test statistic.

---

[21] It is plausible that the prevalences between the evaluation data set and the actual project are the same or almost the same in situations where a classifier is constructed using release k, and then it is used to make predictions for release k+1.

It can also be demonstrated that when *comparing* classifiers, misleading conclusions can be drawn if one uses the proportion correct accuracy as the comparison criterion. Consider the two classifiers represented by examples 4 and 5 in Table 11.

Classifier 4 has absolutely no capability in classifying any of the high risk components correctly. The proportion correct accuracy of both of these classifiers is 0.64 for a prevalence of 0.2. Therefore, on the same data set with the same prevalence, both classifiers look the same in terms of proportion correct accuracy. However, intuitively, the first classifier is quite useless.

Therefore, we can conclude that even for comparing different classifiers, the use of proportion correct accuracy can provide quite misleading results, and should therefore be abandoned.

The problem with using proportion correct accuracy is known in the medical sciences [19][52]. For example, If one is evaluating a diagnostic test in a sample of the population that is at high-risk of having a disease, then prevalence is likely to be higher than if evaluating it in a sample from the general population whereby the prevalence is likely to be lower. The results from both of these evaluation contexts can give dramatically different results as to the accuracy of the test, and certainly one cannot justify applying the test using only the accuracy information from the high risk sample with the intent of using it on the general population.

6.7.3 Correctness

Correctness can be reformulated as follows:

$$Corr = \frac{sp_h}{(1-p_h)(1-f)+sp_h} \qquad \textbf{Eqn. 10}$$

This equation makes clear that correctness is highly dependent on prevalence. Let us consider example 1 in Table 10. We have an appealing correctness of 0.98 on the evaluation data set, but a correctness of only 0.636 on the real project data set, D_R. This is a substantial reduction in correctness that would occur in practice compared to the evaluation results. The relationship between prevalence and correctness is clearly seen in Figure 8. A situation more congruent with current practice is exemplified by example 3. Here we have a seemingly respectable correctness of 0.67 for the evaluation data set, but this drops to a negligible 0.09 for the actual data set. Therefore, from these examples it can be seen that correctness is extremely sensitive to prevalence.

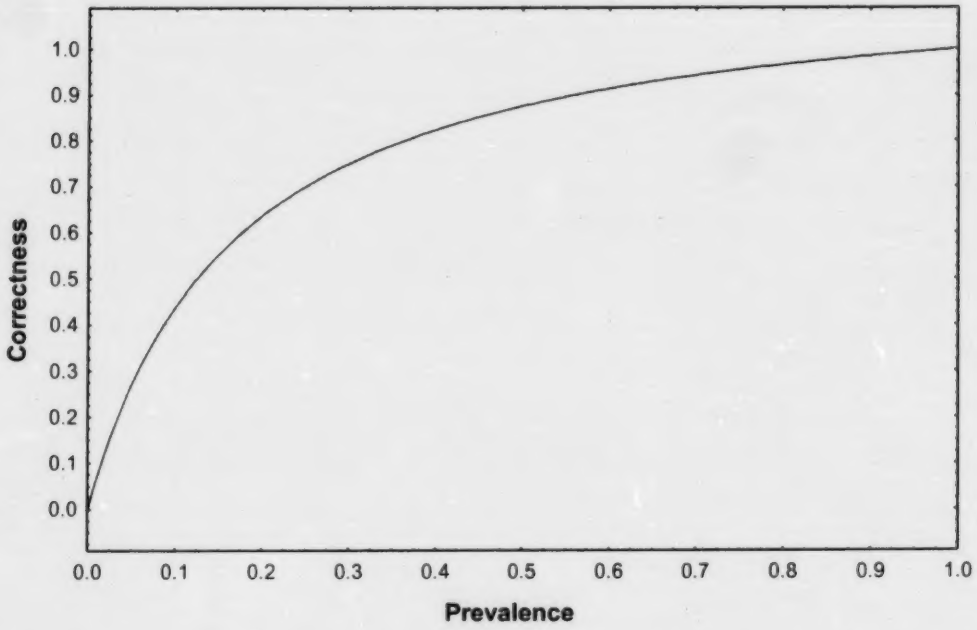

Figure 8: Plot of prevalence versus correctness for s=0.7 and f=0.9 (examples 1 and 2 in Table 10).

6.7.4 The Kappa Coefficient

It is known that the Kappa coefficient has a strong dependence on prevalence [21][51]. The Kappa coefficient can be expressed in terms of prevalence, the sensitivity, and the specificity as follows:

$$\kappa = \frac{2p_h(p_h - 1)(1 - f - s)}{(1 - p_h) - (1 - 2p_h)[f(1 - p_h) + p_h(1 - s)]} \qquad \text{Eqn. 11}$$

It is clear from this equation that Kappa depends strongly on the prevalence value. The problem with this dependence can be seen in Figure 9. At extremes of prevalance Kappa tends to decrease for a fixed sensitivity and specificity. Therefore, if the prevalence during evaluation, p_h, is set at 0.5 then Kappa is approximately 0.6, which would seem to be a high value. However, if during the application of the classifier prevalence, π, is lower than 0.1 then Kappa dips quite fast. Therefore, it is clear that the use of Kappa with the common practice of median splits can lead to rather optimistic conclusions about the value of a classifier.

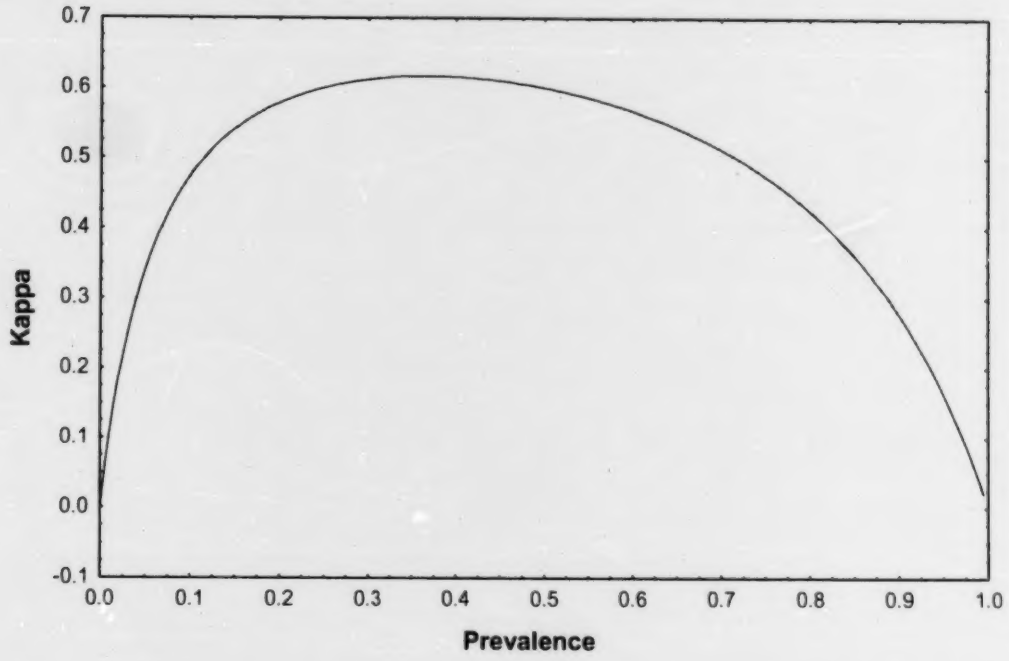

Figure 9: Plot of prevalence versus Kappa for s=0.7 and f=0.9 (examples 1 and 2 in Table 10).

It has been suggested that when one of the marginals is fixed (which is the case with a confusion matrix) that the chance "agreement" should be defined as proportion correct obtained by guessing, in this case with a prior probability of 0.5. However, as noted in [16], this results in a linear transformation of the proportion correct coefficient, and therefore will still suffer from the same problems.

Another suggestion was to evaluate improvement over a "best" a priori strategy exemplified by $\max(p_h, 1-p_h)$ [5]. Therefore, a new Kappa coefficient would be defined as:

$$\kappa_h = \frac{sp_h + f(1-p_h) - \max(p_h, 1-p_h)}{1 - \max(p_h, 1-p_h)} \qquad \textbf{Eqn. 12}$$

In the case where $\max(p_h, 1-p_h) = p_h$, we have:

$$\kappa_h = \frac{p_h(s-1) + f(1-p_h)}{1-p_h} \qquad \textbf{Eqn. 13}$$

and when $\max(p_h, 1-p_h) = 1-p_h$, we have:

$$\kappa_h = \frac{p_h(s+1) + f(1-p_h) - 1}{p_h} \qquad \textbf{Eqn. 14}$$

As can be seen, both of these equations also exhibit a strong dependence on the prevalence, and therefore the results obtained from the evaluation phase are unlikely to generalize to the application phase.

6.7.5 The Information Score

Recognizing the problems with criteria for evaluating classifiers, Kononenko and Bratko [37] propose a new coefficient based on information theory. While this coefficient has not, to our knowledge, been used in software engineering, we include it here to pre-empt its application in a software engineering context. Their coefficient however explicitly takes into account the prevalence p_h. Therefore, given that in software engineering contexts $p_h \neq \pi$, it follows that such a coefficient would not be useful in our context. We illustrate this below.

In an evaluation context, the prior probability of a risk class $C \in \{High, Low\}$ is given by the prevalence p_h. For every observation i in D_h the posterior probability of being in a "High" risk class is denoted by p_i' and of being in a "Low" risk class by $1 - p_i'$. Some classifiers, such as decision trees [12][48] return a subset at the leaves and take the most frequent class as the predicted value. However, each leaf has a distribution and therefore p_i' can take on any value between 0 and 1. Other classifiers, such as k nearest neighbor with k=1, return a single similar observation from the case base, and therefore p_i' can have values of either 0 or 1.

The expected amount of information for classifying a single observation is given by the entropy:

$$E = (p_h \times \log_2(p_h)) + ((1-p_h) \times \log_2(1-p_h)) \qquad \textbf{Eqn. 15}$$

Let us say that the correct risk class of a particular observation is "High". The authors then define three cases:

1. If $p_i' > p_h$ then the probability of being in a "High" class has moved in the right direction.
2. If $p_i' < p_h$ then the probability of being in a "High" risk class has moved in the wrong direction.
3. If $p_i' = p_h$ then this classification contains no information.

An information score I is then defined as shown below.

If $p_i' \geq p_h$ then:

$$I_C(i) = -\log_2(p_h) + \log_2(p_i') \qquad \textbf{Eqn. 16}$$

which is the information necessary to correctly classify an instance into the "High" risk class, minus the remainder of information necessary to correctly classify that instance.

If $p_i' < p_h$ then:

$$I_M(i) = -\log_2(1-p_h) + \log_2(1-p_i') \qquad \textbf{Eqn. 17}$$

which is the information necessary to decide that an instance does not belong to the "High" risk class, minus the remainder of information necessary to make the decision.

The average information score is then given by:

$$I_a = \frac{1}{|D_h|}\sum_{i}^{|D_h|} I(i) \qquad \textbf{Eqn. 18}$$

where $I(i)$ is $I_C(i)$ or $I_M(i)$ depending on the case. The relative information score is defined as:

$$I_r = \left(\frac{I_a}{E}\right)\times 100 \qquad \textbf{Eqn. 19}$$

It is suggested that I_a be used to compare different classifiers on the same data set, and I_r be used to compare classifiers on different data sets.

Now, consider the scenario where we have a perfect classifier that always classifies in the correct class with a probability of 1. Furthermore, assume that our prevalence on the holdout sample, p_h, is 0.5. The average information score obtained during evaluation for this classifier is $I_a = 1$. However, if the actual prevalence $\pi = 0.2$, then we have the actual average information score of $I_a = 0.72$. Therefore, in this particular example, the performance of the classifier was exaggerated during the evaluation phase because the evaluation prevalence was larger than the actual prevalence. By reversing the prevalence values we can get an example of an evaluation that underestimates the performance of a classifier.

6.7.6 Summary

The above exposition makes clear that the chi-square, proportion correct, correctness, Kappa, and information score measures are not adequate in a software engineering context. Furthermore, we have argued that multiple measures, such as Type I and Type II error rates when used together in comparing classifiers have been known to give conflicting results, making it more difficult to draw objective conclusions. Below we propose a single measure of accuracy that alleviates these problems.

6.8 The J Coefficient

To address deficiencies in evaluative measures in the medical sciences, Youdon has proposed the J coefficient[22] [57]. The logic of the J coefficient is described below.

A measure of success of a classifier in identifying low risk components is:

$$\frac{n_{11} - n_{12}}{n_{11} + n_{12}} \qquad \textbf{Eqn. 20}$$

A measure of success of a classifier in identifying high risk components is:

[22] Clearly, as noted in [51], it is plausible to devise a number of evaluative measures that are independent of prevalence. Amongst these, there is no compelling reason to prefer one of the other, and therefore we use the J coefficient due to its obvious relationship to specificity and sensitivity.

$$\frac{n_{22} - n_{21}}{n_{22} + n_{21}}$$ **Eqn. 21**

The J coefficient is defined as the average of these two values:

$$J = \frac{1}{2}\left(\frac{n_{11} - n_{12}}{n_{11} + n_{12}} + \frac{n_{22} - n_{21}}{n_{22} + n_{21}}\right) = s + f - 1$$ **Eqn. 22**

This coefficient has a number of desirable properties. First, it is prevalence independent. This can be illustrated with reference to our examples above. If our classifier has specificity and sensitivity equal to $f = 0.9$ and $s = 0.7$, then its J value is 0.6 irrespective of prevalence in the evaluation sample and in the real project. For our two classifiers in the second example set, the values would be $J_1 = -0.2$ and $J_2 = 0.4$ respectively, which makes intuitive sense since the first classifier uses only one class in its classifications.

The J coefficient can vary from minus one to plus one, with plus one being perfect accuracy and –1 being the worst accuracy. A guessing classifier (i.e., one that guesses High/Low risk with a probability of 0.5) would have a J value of 0. Therefore, J values greater than zero indicate that the classifier is performing better than would be expected from a guessing classifier.

The J coefficient is used in our study to evaluate and compare classifiers on the same data set.

7 Acknowledgements

We wish to thank Anatol Kark for his comments on an earlier version of this paper.

8 References

[1] M. Almeida, H. Lounis, and W. Melo: "An Investigation on the Use of Machine Learned Models for Estimating Correction Costs". In *Proceedings of the 20th International Conference on Software Engineering*, pp. 473-476, 1998.

[2] K-D Althoff, E. Auriol, R. Barletta, and M. Manago: *A Review of Industrial Case-Based Reasoning Tools*. AI Intelligence, 1995.

[3] P. Armitage and G. Berry: *Statistical Methods in Medical Research*. Blackwell Science, 1994.

[4] V. Basili, S. Condon, K. El Emam, R. Hendrick, and W. Melo: "Characterizing and Modeling the Cost of Rework in a Library of Reusable Software Components". In *Proceedings of the 19th International Conference on Software Engineering*, 1997.

[5] R. Brennan and D. Prediger: "Coefficient Kappa: Some Uses, Misuses, and Alternatives". In *Educational and Psychological Measurement*, 41:687-699, 1981.

[6] L. Briand, V. Basili, and W. Thomas: "A Pattern Recognition Approach for Software Engineering Data Analysis". In *IEEE Transactions on Software Engineering*, 18(11):931-942, 1992.

[7] L. Briand, V. Basili, and C. Hetmanski: "Developing Interpretable Models with Optimized Set Reduction for Identifying High-Risk Software Components". In *IEEE Transactions on Software Engineering*, 19(11):1028-1044, 1993.

[8] L. Briand, W. Thomas, and C. Hetmanski: "Modeling and Managing Risk Early in Software Development". In *Proceedings of the International Conference on Software Engineering*, pages 55-65, 1993.

[9] L. Briand, J. Wuest, S. Ikonomovski, and H. Lounis: "A Comprehensive Investigation of Quality Factors in Object-Oriented Designs: An Industrial Case Study". International Software Engineering Research Network technical report ISERN-98-29, 1998.

[10] L. Briand, J. Wuest, S. Ikonomovski, and H. Lounis: "Investigating Quality Factors in Object-Oriented Designs: An Industrial Case Study". In *Proceedings of the International Conference on Software Engineering*, 1999.

[11] L. Briand, J. Wuest, J. Daly, and V. Porter: "Exploring the Relationships Between Design Measures and Software Quality in Object Oriented Systems". To appear in *Journal of Systems and Software*.

[12] L. Breiman, J. Friedman, R. Olshen, and C. Stone: *Classification and Regression Trees*. Wadsworth, 1984.

[13] F. Coallier, J. Mayrand, and B. Lague: "Risk Management in Software Product Procurement". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[14] J. Cohen: "A Coefficient of Agreement for Nominal Scales". In *Educational and Psychological Measurement*, XX(1):37-46, 1960.

[15] J. Cohen: "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Agreement or Partial Credit". In *Psychological Bulletin*, 70:213-220, 1968.

[16] K. El Emam: "The Predictive Validity Criterion for Evaluating Binary Classifiers". In *Proceedings of the 5th International Symposium on Software Metrics*, pages 235-244, IEEE CS Press, 1998.

[17] N. Fenton and N. Ohlsson: "Quantitative Analysis of Faults and Failures in a Complex Software System". To appear in *IEEE Transactions on Software Engineering*.

[18] K. Ganesan, T. Khoshgoftaar, and E. Allen: "Case-based Software Quality Prediction". To appear in the *International Journal of Software Engineering and Knowledge Engineering*, 1999.

[19] L. Gordis: *Epidemiology*. W. B. Saunders Company, 1996.

[20] A. Gray and S. MacDonell: "A Comparison of Techniques for Developing Predictive Models of Software Metrics". In *Information and Software Technology*, 39:425-437, 1997.

[21] W. Grove, N. Andreasen, P. McDonald-Scott, M. Keller, and R. Shapiro: "Reliability Studies of Psychiatric Diagnosis: Theory and Practice". In *Archives of General Psychiatry*, 38:408-413, 1981.

[22] W. Harrison: "Using Software Metrics to Allocate testing Resources". In *Journal of Management Information Systems*, 4(4):93-105, 1988.

[23] D. Hoaglin, F. Mosteller, and J. Tukey (eds.): *Understanding Robust and Exploratory Data Analysis*. Wiley, 1983.

[24] D. Hosmer and S. Lemeshow: *Applied Logistic Regression*. John Wiley & Sons, 1989.

[25] M. Kaaniche and K. Kanoun: "Reliability of a Commercial Telecommunications System". In *Proceedings of the International Symposium on Software Reliability Engineering*, pages 207-212, 1996.

[26] D. Kafura and S. Henry: "Software Quality Based on Interconnectivity". In *Journal of Systems and Software*, 2:121-131, 1981.

[27] L. Kaufman and P. Rousseeuw: *Finding Groups in Data*. John Wiley & Sons, 1990.

[28] T. Khoshgoftaar, E. Allen, J. Hudepohl, S. Aud, and J. Mayrand: "Selecting Software Metrics for a Large Telecommunications System". In *Proceedings of the Fourth Software Engineering Research Forum*, pp. 221-229, 1995.

[29] T. Khoshgoftaar, E. Allen, K. Kalaichelvan, N. Goel, J. Hudepohl, and J. Mayrand: "Detection of Fault-Prone Program Modules in a Very Large Telecommunications System". In *Proceedings of the 6th International Symposium on Software Reliability Engineering*, pp. 24-33, 1995.

[30] T. Khoshgoftaar, E. Allen, K. Kalaichelvan, and N. Goel: "The Impact of Software Evolution and Reuse on Software Quality". In *Empirical Software Engineering: An International Journal*, 1:31-44, 1996.

[31] T. Khoshgoftaar, E. Allen, L. Bullard, R. Halstead, and G. Trio: "A Tree Based Classification Model for Analysis of a Military Software System". In *Proceedings of the IEEE High-Assurance Systems Engineering Workshop*, pages 244-251, 1996.

[32] T. Khoshgoftaar, K. Ganesan, E. Allen, F. Ross, R. Munikoti, N. Goel, and A. Nandi: "Predicting Fault-Prone Modules with Case-Based Reasoning". In *Proceedings of the Eighth International Symposium on Software Reliability Engineering*, pages 27-35, 1997.

[33] T. Khoshgoftaar, E. Allen, and D. Lanning: "An Information Theory-Based Approach to Quantifying the Contribution of a Software Metric". In *Journal of Systems and Software*, 36:103-113, 1997.

[34] T. Khoshgoftaar and E. Allen: "The Impact of Costs of Misclassification on Software Quality Modeling". In *Proceedings of the 4th International Symposium on Software Metrics*, pages 54-62, 1997.

[35] T. Khoshgoftaar, E. Allen, W. Jones, and J. Hudepohl: "Classification Tree Models of Software Quality Over Multiple Releases". To appear in *Proceedings of the International Symposium on Software Reliability Engineering*, 1999.
[36] J. Kolodner: *Case-Based Reasoning*. Morgan Kaufmann Publishers Inc., 1993.
[37] I. Kononeko and I. Bratko: "Information-Based Evaluation Criterion for Classifier's Performance". In *Machine Learning*, 6:67-80, 1991.
[38] F. Lanubile and G. Visaggio: "Evaluating Predictive Quality Models Derived from Software Measures: Lessons Learned". In *Journal of Systems and Software*, 38:225-234, 1997.
[39] J. Mayrand and F. Coallier: "System Acquisition Based on Software Product Assessment". In *Proceedings of the 18th International Conference on Software Engineering*, pp. 210-219, 1996.
[40] T. McCabe: "A Complexity Measure". In *IEEE Transactions on Software Engineering*, 2(4):308-320, 1976.
[41] G. Milligan and M. Cooper: "A Study of Standardization of Variables in Cluster Analysis". In *Journal of Classification*, 5:181-204, 1988.
[42] K-H Moller: "An Empirical Investigation of Software Fault Distribution". In *Proceedings of CSR*, pages 82-90, 1993.
[43] A. Myrvold: "Data Analysis for Software Metrics". In *Journal of Systems and Software*, 12:271-275, 1990.
[44] J. Navlakha: "Measuring the Effect of External and Internal Interface on Software Development". In *Proceedings of the 20th Annual International Conference on the System Sciences*, pp. 127-136, 1987.
[45] N. Ohlsson and H. Alberg: "Predicting Fault-Prone Software Modules in Telephone Switches". In *IEEE Transactions on Software Engineering*, 22(12):886-894, 1996.
[46] A. Porter and R. Selby: "Evaluating Techniques for Generating Metric-Based Classification Trees". In *Journal of Systems and Software*, 12:209-218, 1990.
[47] A. Porter: "Using Measurement-Driven Modeling to Provide Empirical Feedback to Software Developers". In *Journal of Systems and Software*, 20:237-243, 1993.
[48] J. Quinlan: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
[49] N. Schneidewind: "Validating Metrics for Ensuring Space Shuttle Flight Software Quality". In *IEEE Computer*, pages 50-57, August 1994.
[50] N. Schneidewind: "Software Metrics Model for Integrating Quality Control and Prediction". In *Proceedings of the 8th International Symposium on Software Reliability Engineering*, pages 402-415, 1997.
[51] E. Spitznagel and J. Helzer: "A Proposed Solution to the Base Rate Problem in the Kappa Statistic". In *Archives of General Psychiatry*, 42:725-728, 1985.
[52] T. Vecchio: "Predictive Value of a Single Diagnostic Test in Unselected Populations". In *The New England Journal of Medicine*, 274(21):1171-1173, 1966.
[53] W. Venables and B. Ripley: *Modern Applied Statistics with S-Plus*. Springer, 1997.
[54] I. Watson: *Applying Case-Based Reasoning: Techniques for Enterprise Systems*. Morgan Kaufmann Publishers Inc., 1997.
[55] H. Weisberg: *Central Tendency and Variability*. Sage Publications, 1992.
[56] J. Yerushalmy: "Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques". In *Public Health Report*, 62:1432-1449, 1947.
[57] W. Youden: "Index for Rating Diagnostic Tests". In *Cancer*, 3:32-35, 1950.